Web Analytics using Map Reduce

Priyanka Lalwani , Dr. Gaurav Aggarwal, Prof. Abhinav Nirwal

***Abstract*— With the advent of E-Commerce, the rapid growth that has occurred in the generation of huge amount of data is remarkable. The use of Internet and the web to transact business generates inventory which should be updated very quickly to remain competitive in the current digital market. Analyzing this web data is very important for companies to predict their customer behavior. Clickstream data is an information trail a user leaves behind while visiting a website. An extremely large volume of data can be collected from the user clicks which can be analyzed and interpreted through Hadoop to generate reports for e business companies. The primary focus of the paper is to prepare web based analysis system which will depict trends based on the users browsing mode using Hadoop MapReduce.**

***Index Terms*— Clickstream, e business, Hadoop, Map Reduce**

I. INTRODUCTION

We are living in an ag*e of Data.* It is not possible to measure the electronic data generated daily but a rough estimate puts the digital data at a size of around 4.4 zettabytes and is forecasted to increase ten times in next 5-8 years. The biggest challenge for the e-commerce companies is to use this huge data to analyze customer click –on and purchase behavior so that they can devise the appropriate strategies based on the customer behavior to enhance the product sales.

Clickstream data is an information trail a user leaves behind while visiting a website. It is typically captured in semi-structured website log files. These website log files contain data elements such as a date and time stamp, the visitor's IP address, the destination URLs of the pages visited, and a user ID that uniquely identifies the website visitor. Mining these log files will be of great use for the e-commerce companies to get the information about customers.

II. APPLICATIONS OF CLICK STREAM DATA

Originally, Hadoop was used to store and process massive volume of clickstream data. But now the e- commerce industry is using Hadoop and Cloudera to filter, refine and analyze clickstream data to improve digital marketing.

The e-commerce companies are now using this analysis for *Path Optimization* which finds the most efficient path for a site visitor to buy a product. One more important application of Clickstream data is Association Analysis which means what products the customer tend to buy together. Using

association analysis the companies can also get the information of the next product the customer would like to buy. The companies can also enhance the user experience by fixing the allocation of website resources.

III. MAP REDUCE

Map Reduce is a data processing model .It is used to handle unstructured data like xml files and structured data like data stored in database in row and column format. Map Reduce can be used to solve complex business logic. Data processing in MapReduce is done in two phases the map phase and the reduce phase. The input and output of each of these phases has key-value pairs. The program has two functions, the map function and the reduce function. A map function prepares the data so that the reduce function works on it to get the desired output.

A MapReduce *job* consists of the input data, the MapReduce program, and configuration information. Hadoop runs the job by dividing it into *tasks*, of which there are two types: *map tasks* and *reduce tasks*. The tasks are scheduled by the job scheduler using YARN (Yet another Resource Negotiator) and run on nodes in the cluster. The system is designed in a fault tolerant manner i.e in case of a failure of a node, the task is being taken by other node in the cluster.

Hadoop divides the input into a fixed-size pieces called *input splits*, or just *splits*. Hadoop creates one mapper task for each of the split, which runs the user-defined map function for each *record* in the split. Dividing the task into many splits means the time taken to process each split is small compared to the time to process the whole input. These splits are processed in parallel so that the processing is better load balanced when the splits are small, since a faster machine will be able to process proportionally more splits over the course of the job than a slower machine. Even if the machines are identical, failed processes or other jobs running concurrently make load balancing desirable, and the quality of the load balancing increases as the splits become small.

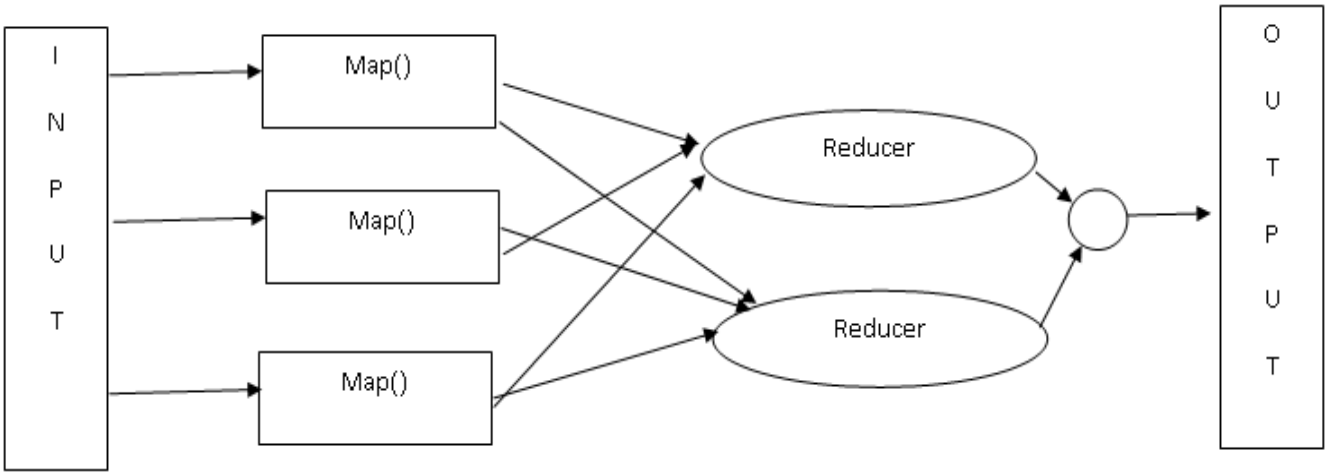
But if splits are too small, the overhead of managing the splits and map task creation begins to dominate the total job execution time. For most jobs, a good split size tends to be the size of an HDFS block, which is 128 MB by default, although this can be changed by the user.

Hadoop does its best to run the map task on a node where the input data resides in HDFS, because it doesn't use valuable cluster bandwidth. This is called the *data locality optimization*. Sometimes, however, all the nodes hosting the HDFS block replicas for a map task's input split are running other map tasks, so the job scheduler will look for a free map slot on a node in the same rack as one of the blocks.

The output from a map function is an intermediate output as this output will be consumed by the reducer to get the final output. The input to a reducer is fed by all mappers. The output of a reducer is stored in HDFS (Hadoop Distributed File System). The data flow between the map and the reducer is called as "Shuffle". Reducer has three primary phases. In the first phase which is called as Shuffle in which

the Reducer copies the sorted output from each Mapper using HTTP across the network. In the second phase called as "sort", the reducer sorts the input by keys. The above two phases run simultaneously which means while the reducer fetches the input it performs the sort and merges the data. A secondary sort is also applied by the reducer in the second phase by defining a grouping comparator so as to know which keys and values are sent in the same call to reduce. In the third phase of the reduce function, a reduce method is called for each key and the output is written to a RecordWriter

IV. MapReduce Framework



V. Conclusion

This paper thus provides insights to process and analyze web log data using MapReduce. Mapreduce can be used to process a large amount of data and to filter the desired output data. In order to optimize your website and convert more visits into sales and revenue.

- Analyze the clickstream data by location
- Filter the data by product category.
- Graph the website user data by age and gender
- Pick a target customer segment
- Identify a few web pages with the highest bounce rates

References

[1] What is big data: - IBM?
[2] "Why Big Data is a must in E-Commerce", Guest post by Jerry Jao, CEO of Retention Science.
http://www.bigdatalandscape.com/news/why-big-data-is-amust-in-ecommerce
[3] Tom White, (2009) "Hadoop: The Definitive Guide. O'Reilly", Scbastopol, California.
[4] Apache-Hadoop, http://Hadoop.apache.org
[5] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai, "ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011